High-Dimensional Multivariate Bayesian Linear Regression with Shrinkage Priors

Ray Bai

Department of Statistics, University of Florida

Joint work with Dr. Malay Ghosh

March 20, 2018

Overview

Simultaneous Prediction and Estimation

There are many scenarios where we would want to simultaneously predict q continuous response variables $y_1, ..., y_q$:

- **Longitudinal data**: The q response variables represent measurements at q consecutive time points.
  - mRNA levels at different time points
  - children's heights at different ages of development
  - CD4 cell counts over time for HIV/AIDS patients
- **The data have a group structure**: The q response variables represent a "group."
  - In genomics, genes within the same pathway often act together in regulating a biological system.

Multivariate Linear Regression

Consider the multivariate linear regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where $\mathbf{Y} = (y_1, ..., y_q)$ is an $n \times q$ response matrix of n samples and q response variables, $\mathbf{X}$ is an $n \times p$ matrix of n samples and p covariates, $\mathbf{B} \in \mathbb{R}^{p \times q}$ is the coefficient matrix, and $\mathbf{E} = (\varepsilon_1, ..., \varepsilon_n)^T$ is an $n \times q$ noise matrix, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N_q(\mathbf{0}, \mathbf{\Sigma}), i = 1, ..., n$.

Throughout, we assume that $\mathbf{X}$ is centered, so there is no intercept term.

Multivariate Linear Regression

For the multivariate linear regression model,

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\mathbf{B}_{p\times q} + \mathbf{E}_{n\times q},$$

where $\mathbf{E} = (\varepsilon_1, ..., \varepsilon_n)^T$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N_q(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, ..., n$,

- $\mathbf{\Sigma}$ represents the covariance structure of the q response variables.
- We wish to estimate the coefficient matrix $\mathbf{B}$.
- Model selection from the p covariates is also often desired. This can be done using multivariate generalizations of AIC, BIC, or Mallow's C_p.

Multivariate Linear Regression

For the multivariate linear regression model, the usual maximum likelihood estimator (MLE) is the ordinary least squares estimator,

$$\widehat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

- The MLE is only unique if $p \leq n$.
- It is well-known that the MLE is an *inconsistent* estimator of $\mathbf{B}$ if $p/n \to c, c > 0$.
- Variable selection using AIC, BIC, and Mallow's C_p is infeasible for large p, since it requires searching over a model space of 2^p models.

High-Dimensional Multivariate Linear Regression

To handle cases where p is large (including the $p > n$ regime), frequentists typically use penalized regression (e.g. Li et al. (2015), Vincent and HAnsen (2014), Wilms and Croux (2017)):

$$\min_{\mathbf{B}} ||\mathbf{Y} - \mathbf{XB}||_2^2 + \lambda \sum_{i=1}^{p} ||\mathbf{b}_i||_2,$$

where $\mathbf{b}_i$ represents the ith row of $\mathbf{B}$ and $\lambda > 0$ is a tuning parameter.

- The group lasso penalty, $||\cdot||_2$, shrinks entire rows of $\mathbf{B}$ to exactly $\mathbf{0}$, leading to a **sparse** estimate of $\mathbf{B}$ and facilitating variable selection from the p estimators.
- We can use adaptive group lasso penalty to avoid overshrinkage of $\mathbf{b}_i, i = 1, ..., p$.

Bayesian High-Dimensional Multivariate Linear Regression

The Bayesian approach is to put a prior distribution on $\mathbf{B}$, $\pi(\mathbf{B})$. That is, given the model, $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ and data $(\mathbf{X}, \mathbf{Y})$, we have

$$\pi(\mathbf{B}|\mathbf{Y}) \propto f(\mathbf{Y}|\mathbf{X}, \mathbf{B})\pi(\mathbf{B}).$$

Inference can be conducted through the posterior, $\pi(\mathbf{B}|\mathbf{Y})$.

Bayesian High-Dimensional Multivariate Linear Regression

To achieve sparsity and variable selection, a common approach is to place spike-and-slab priors on the rows of $\mathbf{B}$ (e.g. Brown et al. (1998), Liquet et al. (2017)):

$$\mathbf{b}_i^T \overset{i.i.d.}{\sim} (1-p)\delta_{\{\mathbf{0}\}} + p\mathcal{N}_q(\mathbf{0}, \tau^2\mathbf{V}), \qquad i = 1, ..., p.$$

- $\delta_{\{\mathbf{0}\}}$ represents a point mass at $\mathbf{0} \in \mathbb{R}^q$, and $\mathbf{V}$ is a $q \times q$ symmetric positive definite matrix.
- τ^2 can be treated as a tuning parameter, or a prior can be placed on τ^2.
- A prior can also be placed on p so that the model adapts to the underlying sparsity. Usually, we put a Beta prior on p.

Bayesian High-Dimensional Multivariate Linear Regression

For the spike-and-slab approach,

$$\begin{gathered}\mathbf{b}_i^T \overset{i.i.d.}{\sim} (1-p)\delta_{\{\mathbf{0}\}} + p\mathcal{N}_q(\mathbf{0}, \tau^2\mathbf{V}), \qquad i = 1, ..., p, \\ \tau^2 \sim \mu(\tau^2), \\ p \sim \mathcal{B}(a, b),\end{gathered}$$

- Taking the posterior median will give a point estimate of $\mathbf{B}$ with rows equal to $\mathbf{0}^T$, thus recovering a sparse estimate of $\mathbf{B}$ and facilitating variable selection.
- Due to the point mass at $\mathbf{0}$, this model can be very, very slow for large p.

Bayesian High-Dimensional Multivariate Linear Regression

Due to the computational inefficiency of discontinuous priors, it is often desirable to put a *continuous* prior on the parameters of interest.

For the multivariate linear regression model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

our aim to estimate $\mathbf{B}$.

- This requires putting a prior density on a $p \times q$ matrix.
- A popular continuous prior to place on $\mathbf{B}$ is the **matrix-normal prior**.

The Matrix-Normal Prior

Definition

A random matrix $\mathbf{X}$ is said to have the *matrix-normal density* if $\mathbf{X}$ has the density function (on the space $\mathbb{R}^{a\times b}$):

$$f(\mathbf{X}) = \frac{|\mathbf{U}|^{-b/2}|\mathbf{V}|^{-a/2}}{(2\pi)^{ab/2}} e^{-\frac{1}{2}\mathrm{tr}[\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})^{\mathsf{T}}]},$$

where $\mathbf{M} \in \mathbb{R}^{a\times b}$, and $\mathbf{U}$ and $\mathbf{V}$ are positive semi-definite matrices of dimension $a \times a$ and $b \times b$ respectively. If $\mathbf{X}$ is distributed as a matrix-normal distribution with pdf above, we write $\mathbf{X} \sim MN_{a\times b}(\mathbf{M}, \mathbf{U}, \mathbf{V})$.

Multivariate Bayesian Model with Shrinkage Priors (MBSP)

By adding an *additional* layer in the Bayesian hierarchy, we can obtain a *row-sparse* estimate of $\mathbf{B}$. This row-sparse estimate also facilitates variable selection from the p variables. Our model is specified as follows:

$$\begin{aligned} \mathbf{Y}|\mathbf{X},\mathbf{B},\mathbf{\Sigma} &\sim MN_{n\times q}(\mathbf{XB},\mathbf{I}_n,\mathbf{\Sigma}), \\ \mathbf{B}|\xi_1,...,\xi_p,\mathbf{\Sigma} &\sim MN_{p\times q}(\mathbf{O},\tau\,\mathrm{diag}(\xi_1,...,\xi_p),\mathbf{\Sigma}), \\ \xi_i &\overset{\text{ind}}{\sim} \pi(\xi_i), i=1,...,p, \end{aligned}$$

where $\tau > 0$ is a tuning parameter, and $\pi(\xi_i)$ is a *polynomial-tailed* prior density of the form,

$$\pi(\xi_i) = K(\xi_i)^{-a-1}L(\xi_i),$$

where $K > 0$ is the constant of proportionality, a is positive real number, and L is a a positive measurable, non-constant, slowly varying function over $(0,\infty)$.

Examples of Polynomial-Tailed Priors

Prior	$\pi(\xi_i)/C$	$L(\xi_i)$
Student's t	$\xi_i^{-a-1}\exp(-a/\xi_i)$	$\exp -a/\xi_i$
Horseshoe	$\xi_i^{-1/2}(1+\xi_i)^{-1}$	$\xi_i^a/(1+\xi_i)$
Horseshoe+	$\xi_i^{-1/2}(\xi_i-1)^{-1}\log(\xi_i)$	$\xi_i^a(\xi_i-1)^{-1}\log(\xi_i)$
NEG	$(1+\xi_i)^{-1-a}$	$\{\xi_i/(1+\xi_i)\}^{a+1}$
TPBN	$\xi_i^{u-1}(1+\xi_i)^{-a-u}$	$\{\xi_i/(1+\xi_i)\}^{a+u}$
GDP	$\int_0^\infty \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2\xi_i}{2}\right)\lambda^{2a-1}\exp(-\eta\lambda)d\lambda$	$\int_0^\infty t^a\exp(-t-\eta\sqrt{2t/\xi_i})dt$
HIB	$\xi_i^{u-1}(1+\xi_i)^{-(a+u)}\exp\left\{-\frac{s}{1+\xi_i}\right\}$ $\times\left\{\phi^2+\frac{1-\phi^2}{1+\xi_i}\right\}^{-1}$	$\{\xi_i/(1+\xi_i)\}^{a+u}$ $\times\exp\left\{-\frac{s}{1+\xi_i}\right\}\left\{\phi^2+\frac{1-\phi^2}{1+\xi_i}\right\}^{-1}$

Table: Polynomial-tailed priors, their respective prior densities for $\pi(\xi_i)$ up to normalizing constant C, and the slowly-varying component $L(\xi_i)$.

Sparse Estimation of $\mathbf{B}$: Examples

If $\pi(\xi_j) \overset{\text{ind}}{\sim} \text{Inverse-Gamma}(\alpha_j, \frac{\gamma_j}{2})$, then the marginal density for $\mathbf{B}$, $\pi(\mathbf{B})$, under the MBSP model is proportional to

$$\prod_{j=1}^{p} \left(||\mathbf{b}_j (\tau \mathbf{\Sigma})^{-1/2}||_2^2 + \gamma_j \right)^{-(\alpha_j + \frac{q}{2})},$$

which corresponds to a multivariate t-distribution. Here $\mathbf{b}_j$ denotes the jth row of $\mathbf{B}$.

Sparse Estimation of $\mathbf{B}$: Examples

If $\pi(\xi_j) \propto \xi_j^{q/2-1}(1+\xi_j)^{-1}$, then the joint density $\pi(\mathbf{B}, \xi_1, ..., \xi_p)$ under the MBSP model is proportional to

$$\prod_{j=1}^{p} \xi_j^{-1}(1+\xi_j)^{-1} e^{-\frac{1}{2\xi_j}||\mathbf{b}_j(\tau\mathbf{\Sigma})^{-1/2}||_2^2},$$

and integrating out the ξ_j's gives a multivariate horseshoe density function.

Notation

- For any two sequences of positive real numbers $\{a_n\}$ and $\{b_n\}$ with $b_n \neq 0$,
  - $a_n = O(b_n)$ if $\left|\frac{a_n}{b_n}\right| \leq M$ for all n, for some positive real number M independent of n
  - $a_n = o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. Therefore, $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$.
- For a vector $v \in \mathbb{R}^n$, $||v||_2 := \sqrt{\sum_{i=1}^n v_i^2}$ denote the ℓ_2 norm.
- For a matrix $\mathbf{A} \in \mathbb{R}^{a\times b}$ with entries a_{ij}, $||\mathbf{A}||_F := \sqrt{\mathrm{tr}(\mathbf{A}^T\mathbf{A})}$ $= \sqrt{\sum_{i=1}^a \sum_{j=1}^b a_{ij}^2}$ denotes the Frobenius norm of $\mathbf{A}$.
- For a symmetric matrix $\mathbf{A}$, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ respectively.

Posterior Consistency

Suppose that the data is generated from a true model,

$$\mathbf{Y}_n = \mathbf{X}\mathbf{B}_0 + \mathbf{E}_n,$$

where $\mathbf{Y}_n := (\mathbf{Y}_{n,1}, ..., \mathbf{Y}_{n,q})$ and $\mathbf{E}_n \sim MN_{n\times q}(\mathbf{O}, \mathbf{I}_n, \mathbf{\Sigma})$.

Letting $\mathbb{P}_0$ denote the probability measure underlying the true model above, we define the following notion of posterior consistency:

Definition

(strong posterior consistency) Let $\mathcal{B}_n = \{\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon\}$, where $\varepsilon > 0$. The sequence of posterior distributions of $\mathbf{B}_n$ under prior $\pi_n(\mathbf{B}_n)$ is said to be *strongly* consistent under the true model if, for any $\varepsilon > 0$,

$$\Pi_n(\mathcal{B}_n|Y_n) = \Pi_n(||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon|\mathbf{Y}_n) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty.$$

Sufficient Conditions for Posterior Consistency

For our theoretical analysis, we assume that $q < n$ is fixed and $\mathbf{\Sigma}$ is known.

- In practice, $\mathbf{\Sigma}$ is often unknown and can be estimated from the data using an Inverse Wishart prior on $\mathbf{\Sigma}$ or by obtaining a separate estimate $\widehat{\mathbf{\Sigma}}$ (e.g. the MLE) and plugging $\widehat{\mathbf{\Sigma}}$ into our model as an empirical Bayes estimate.

Theory is developed separately for:

- $p_n = o(n)$ (low-dimensional setting)
- $p_n \geq O(n)$ (ultrahigh-dimensional setting)

Regularity Conditions for the Low-Dimensional Case

(A1) $p_n = o(n)$ and $p_n \leq n$ for all $n \geq 1$.

(A2) There exist constants c_1, c_2 so that

$$0 < c_1 < \limsup_{n\to\infty} \lambda_{\min}\left(\frac{\mathbf{X}_n^T\mathbf{X}_n}{n}\right) \leq \limsup_{n\to\infty} \lambda_{\max}\left(\frac{\mathbf{X}_n^T\mathbf{X}_n}{n}\right) < c_2 < \infty.$$

(A3) There exist constants d_1 and d_2 so that

$$0 < d_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < d_2 < \infty.$$

Sufficient Conditions for Posterior Consistency When $p = o(n)$

Theorem

Assume that conditions (A1)-(A3) hold. Then the posterior of $\mathbf{B}_n$ *under any prior* $\pi_n(\mathbf{B}_n)$ *is strongly consistent. That is, for any* $\varepsilon > 0$,

$$\Pi_n(\mathcal{B}_n|\mathbf{Y}_n) = \Pi_n(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon|\mathbf{Y}_n) \to 0 \ \ \mathbb{P}_0 \ a.s. \ as \ n \to \infty$$

if

$$\Pi_n\left(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F < \frac{\Delta}{n^{\rho/2}}\right) > \exp(-kn)$$

for all $0 < \Delta < \frac{\varepsilon^2 c_1 d_1^{1/2}}{48 c_2^{1/2} d_2}$ *and* $0 < k < \frac{\varepsilon^2 c_1}{32 d_2} - \frac{3\Delta c_2^{1/2}}{2 d_1^{1/2}}$, *where* $\rho > 0$.

This theorem applies to **any** prior on $\mathbf{B}_n$. Provided the prior satisfies the above condition and $p = o(n)$, the posterior is strongly consistent.

The MBSP Model

Recall the MBSP model:

$$\begin{aligned} \mathbf{Y}|\mathbf{X},\mathbf{B},\mathbf{\Sigma} &\sim MN_{n\times q}(\mathbf{XB},\mathbf{I}_n,\mathbf{\Sigma}), \\ \mathbf{B}|\xi_1,...,\xi_p,\mathbf{\Sigma} &\sim MN_{p\times q}(\mathbf{O},\tau\,\mathrm{diag}(\xi_1,...,\xi_{p_n}),\mathbf{\Sigma}), \\ \xi_i &\overset{\text{ind}}{\sim} \pi(\xi_i), i=1,...,p_n, \end{aligned}$$

where $\tau_n > 0$ and $\pi(\xi_i)$ is a polynomial-tailed density of the form,

$$\pi(\xi_i) = K(\xi_i)^{-a-1}L(\xi_i),$$

To achieve posterior consistency, we require mild conditions on the slowly varying component $L(\cdot)$, $\tau_n > 0$, and the true unknown coefficients matrix $\mathbf{B}_0$.

Additional Assumptions under the MBSP Model

(i) For the slowly varying function $L(t)$ in the priors for $\xi_i, 1 \leq i \leq p_n$, $\lim_{t\to\infty} L(t) \in (0, \infty)$. That is, there exists $c_0 (> 0)$ such that $L(t) \geq c_0$ for all $t \geq t_0$, for some t_0 which depends on both L and c_0.

(ii) There exists $M > 0$ so that $\sup_{j,k} |b^0_{jk}| \leq M < \infty$ for all n, i.e. the maximum entry in $\mathbf{B}_0$ is uniformly bounded above in absolute value.

(iii) $0 < \tau_n < 1$ for all n, and $\tau_n = o\left(\frac{1}{p_n n^\rho}\right)$ for some $\rho > 0$.

Posterior Consistency of MBSP (low-dimensional case)

Theorem

Suppose that we have the MBSP model with polynomial-tailed priors for $\xi_1, ..., \xi_p$. *Provided that Assumptions (A1)-(A3) and (i)-(iii) hold, our model achieves strong posterior consistency. That is, for any* $\varepsilon > 0$,

$$\Pi_n(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon | \mathbf{Y}_n) \to 0 \ \ \mathbb{P}_0 \ a.s. \ as \ n \to 0.$$

Ultrahigh-Dimensional Case

We have shown that the MBSP model achieves posterior consistency under mild conditions if $p_n = o(n)$.

What if $p_n > n$ and $p_n \geq O(n)$?

It turns out that with some additional regularity conditions on the model size and the design matrix, we *can* achieve posterior consistency in this ultrahigh-dimensional setting!

Regularity Conditions for the Ultrahigh-dimensional Case

(B1) $p_n > n$ for all $n \geq 1$, and $\log(p_n) = O(n^d)$ for some $0 < d < 1$.

(B2) The rank of $\mathbf{X}_n$ is n.

(B3) Let $\mathcal{J}$ denote a set of indices, where $\mathcal{J} \subset \{1, ..., p_n\}$ such that $|\mathcal{J}| \leq n$. Let $\mathbf{X}_{\mathcal{J}}$ denote the submatrix of $\mathbf{X}$ that contains the columns with indices in $\mathcal{J}$. For any such set $\mathcal{J}$, there exists a finite constant $\widetilde{c}_1 (> 0)$ so that $\liminf_{n\to\infty} \lambda_{\min} \left(\frac{\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}}}{n} \right) \geq \widetilde{c}_1$.

(B4) There is finite constant $\widetilde{c}_2 (> 0)$ so that

$$\limsup_{n\to\infty} \lambda_{\max} \left(\frac{\mathbf{X}_n^T \mathbf{X}_n}{n} \right) \leq \widetilde{c}_2 < \infty.$$

(B5) There exist constants d_1 and d_2 so that

$$0 < d_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < d_2 < \infty.$$

(B6) The true model $S^* \subset \{1, ..., p_n\}$ is nonempty for all n and $s^* = |S^*| = o(n/log(p_n))$.

Sufficient Conditions for Posterior Consistency When $\log p = o(n)$

Theorem

Assume that conditions B1-B6 hold. Then the posterior of $\mathbf{B}_n$ *under any prior* $\pi_n(\mathbf{B}_n)$ *is strongly consistent. That is, for any* $\varepsilon > 0$,

$$\Pi_n(\mathcal{B}_n|\mathbf{Y}_n) = \Pi_n(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon|\mathbf{Y}_n) \to 0 \ \ \mathbb{P}_0 \ a.s. \ as \ n \to \infty$$

if

$$\Pi_n\left(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F < \frac{\Delta}{n^{\rho/2}}\right) > \exp(-kn)$$

for all $0 < \widetilde{\Delta} < \frac{\varepsilon^2 \widetilde{c}_1 d_1^{1/2}}{48 \widetilde{c}_2^{1/2} d_2}$ *and* $0 < k < \frac{\varepsilon^2 \widetilde{c}_1}{32 d_2} - \frac{3\widetilde{\Delta} \widetilde{c}_2^{1/2}}{2 d_1^{1/2}}$, *where* $\rho > 0$.

This theorem applies to **any** prior on $\mathbf{B}_n$. Provided the prior satisfies the above condition and $\log p = o(n)$, the posterior is strongly consistent.

The MBSP Model

Recall the MBSP model:

$$\begin{aligned} \mathbf{Y}|\mathbf{X},\mathbf{B},\mathbf{\Sigma} &\sim MN_{n\times q}(\mathbf{XB},\mathbf{I}_n,\mathbf{\Sigma}), \\ \mathbf{B}|\xi_1,...,\xi_p,\mathbf{\Sigma} &\sim MN_{p\times q}(\mathbf{O},\tau\,\mathrm{diag}(\xi_1,...,\xi_{p_n}),\mathbf{\Sigma}), \\ \xi_i &\overset{\text{ind}}{\sim} \pi(\xi_i), i=1,...,p_n, \end{aligned}$$

where $\tau_n > 0$ and $\pi(\xi_i)$ is a polynomial-tailed density of the form,

$$\pi(\xi_i) = K(\xi_i)^{-a-1}L(\xi_i),$$

To achieve posterior consistency, we require mild conditions on the slowly varying component $L(\cdot)$, $\tau_n > 0$, and the true unknown coefficients matrix $\mathbf{B}_0$.

Additional Assumptions under the MBSP Model

(i) For the slowly varying function $L(t)$ in the priors for $\xi_i, 1 \leq i \leq p_n$, $\lim_{t\to\infty} L(t) \in (0, \infty)$. That is, there exists $c_0(> 0)$ such that $L(t) \geq c_0$ for all $t \geq t_0$, for some t_0 which depends on both L and c_0.

(ii) There exists $M > 0$ so that $\sup_{j,k} |b^0_{jk}| \leq M < \infty$ for all n, i.e. the maximum entry in $\mathbf{B}_0$ is uniformly bounded above in absolute value.

(iii) $0 < \tau_n < 1$ for all n, and $\tau_n = o\left(\frac{1}{p_n n^\rho}\right)$ for some $\rho > 0$.

- Note that these are the **same** conditions as in the low-dimensional setting!
- The same rate for τ_n works for both low-dimensional and high-dimensional cases.

Posterior Consistency of MBSP (ultrahigh-dimensional case)

Theorem

Suppose that we have the MBSP model with polynomial-tailed priors for $\xi_1, ..., \xi_p$. *Provided that Assumptions (B1)-(B6) and (i)-(iii) hold, our model achieves strong posterior consistency. That is, for any* $\varepsilon > 0$,

$$\Pi_n(\mathbf{B}_n : ||\mathbf{B}_n - \mathbf{B}_0||_F > \varepsilon | \mathbf{Y}_n) \to 0 \ \ \mathbb{P}_0 \ a.s. \ as \ n \to 0.$$

Three Parameter Beta Normal (TPBN) Family

A random variable y said to follow the three parameter beta density, denoted as $TPB(u, a, \tau)$, if

$$\pi(y) = \frac{\Gamma(u+a)}{\Gamma(u)\Gamma(a)} \tau^a y^{a-1} (1-y)^{u-1} \left\{1-(1-\tau)y\right\}^{-(u+a)}.$$

In univariate regression, a global-local shrinkage prior of the form

$$\begin{aligned} \beta_i|\tau, \xi_i &\overset{ind}{\sim} N(0, \tau\xi_i), \quad i = 1, ..., n, \\ \pi(\xi_i) &\overset{ind}{\sim} \frac{\Gamma(u+a)}{\Gamma(u)\Gamma(a)} \xi_i^{u-1} (1+\xi_i)^{-(u+a)}, \quad i = 1, ..., n, \end{aligned}$$

may therefore be represented alternatively as

$$\begin{aligned} \beta_i|\nu_i &\overset{\text{ind}}{\sim} N(0, \nu_i^{-1} - 1), \\ \nu_i &\overset{\text{ind}}{\sim} TPB(u, a, \tau). \end{aligned}$$

Three Parameter Beta Normal (TPBN) Family

After integrating out ν_i in

$$\begin{aligned} \beta_i|\nu_i &\overset{\text{ind}}{\sim} N(0, \nu_i^{-1} - 1), \\ \nu_i &\overset{\text{ind}}{\sim} TPB(u, a, \tau), \end{aligned}$$

the marginal prior for β_i is said to belong to the three parameter beta normal (TPBN) family.

Special cases of the TPBN family include:

- the horseshoe prior ($u = 0.5, a = 0.5$),
- the Strawderman-Berger prior ($u = 1, a = 0.5$),
- the normal-exponential-gamma (NEG) prior ($u = 1, a > 0$).

Three Parameter Beta Normal (TPBN) Model

By Proposition 1 of Armagan et al. (2011), the TPBN prior can also be written as a hierarchical mixture of two Gamma distributions,

$$\beta_i|\psi_i \sim N(0,\psi_i), \quad \psi_i|\zeta_i \sim \mathcal{G}(u,\zeta_i), \quad \zeta_i \sim \mathcal{G}(a,\tau),$$

where $\psi_i = \xi_i\tau$.

Using the TPBN family as our chosen prior and placing a conjugate prior on $\mathbf{\Sigma}$, we can construct a specific variant of the MBSP model which we call the **MBSP-TPBN model**.

MBSP-TPBN Model

Reparametrizing $\psi_i = \tau\xi_i, i = 1, ..., p$, we have:

$$\begin{aligned}
\mathbf{Y}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma} &\sim MN_{n\times q}(\mathbf{XB}, \mathbf{I}_n, \mathbf{\Sigma}), \\
\mathbf{B}|\psi_1, ..., \psi_p, \mathbf{\Sigma} &\sim MN_{p\times q}(\mathbf{O}, \text{diag}(\psi_1, ..., \psi_p), \mathbf{\Sigma}), \\
\psi_i|\zeta_i &\overset{\text{ind}}{\sim} \mathcal{G}(u, \zeta_i), i = 1, ..., p, \\
\zeta_i &\overset{\text{i.i.d.}}{\sim} \mathcal{G}(a, \tau), i = 1, ..., p, \\
\mathbf{\Sigma} &\sim \mathcal{IW}(d, k\mathbf{I}_q),
\end{aligned}$$

The MBSP-TPBN model admits a Gibbs sampler.

Variable Selection

Although the MBSP model and the MBSP-TPBN model produce robust estimates for $\mathbf{B}$, they do not produce exact zeros.

In order to use the MBSP model for variable selection, we recommend looking at the 95% credible intervals for each entry b_{ij} in row i and column j.

- If the credible intervals for every single entry in row $i, 1 \leq i \leq p$, contain zero, then we classify predictor i as an irrelevant predictor.
- If at least one credible interval in row $i, 1 \leq i \leq p$ does not contain zero, then we classify i as an active predictor.

Simulation Study

For our simulation study, we implement the MBSP-TPBN model with the horseshoe prior ($a = u = 0.5$), one of the most popular polynomial priors.

We also set:

- $\tau = \frac{1}{p\sqrt{n \log n}}$
- $d = 3$
- $k =$ variance of residuals, $\mathbf{Y} - \mathbf{X}\mathbf{B}^{(0)}$, where $\mathbf{B}^{(0)}$ is the initial guess in the Gibbs sampler (taken as a ridge estimator).

Simulation Study

Our primary interest is in the $p > n$ case. We consider three different simulation settings with varying levels of sparsity:

- Experiment 1 ($p > n$): $n = 50, p = 200, q = 5$. 20 of the predictors are randomly picked as active (sparse model).
- Experiment 2 ($p > n$): $n = 60, p = 100, q = 6$. 40 of the predictors are randomly picked as active (dense model).
- Experiment 3 ($p \gg n$): $n = 100, p = 500, q = 3$. 10 of the predictors are randomly picked as active (ultra-sparse model).

Simulation Study Metrics

As our point estimate for $\mathbf{B}$, we take the posterior median $\widehat{\mathbf{B}} = (\widehat{B}_{ij})_{p\times q}$. We also perform variable selection by inspecting the 95% credible intervals.

We compute the following metrics, averaged across 100 replications:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{est}} &= 100 \times ||\widehat{\mathbf{B}} - \mathbf{B}||_F^2/(pq), \\
\mathrm{MSE}_{\mathrm{pred}} &= 100 \times ||\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}||_F^2/(nq), \\
\mathrm{FDR} &= \mathrm{FP} \ / \ (\mathrm{TP} + \mathrm{FP}), \\
\mathrm{FNR} &= \mathrm{FN} \ / \ (\mathrm{TN} + \mathrm{FN}), \\
\mathrm{MP} &= (\mathrm{FP} + \mathrm{FN})/(pq),
\end{aligned}$$

where FP, TP, FN, and TN denote the number of false positives, true positives, false negatives, and true negatives respectively.

Simulation Study

Experiment 1: $n = 50, p = 200, q = 5$. 20 active predictors

Method	$\mathbf{MSE_{est}}$	$\mathbf{MSE_{pred}}$	**FDR**	**FNR**	**MP**
MBSP	**1.36**	**117.52**	**0.0117**	0	**0.0013**
MBGL-SS	57.25	694.81	0.858	0.02	0.619
LSGL	8.65	169.30	0.788	0	0.374
SRRR	17.46	161.70	0.698	0	0.307

Experiment 2: $n = 60, p = 100, q = 6$. 40 active predictors

Method	$\mathbf{MSE_{est}}$	$\mathbf{MSE_{pred}}$	**FDR**	**FNR**	**MP**
MBSP	**10.969**	**172.84**	**0.0249**	0	**0.0107**
MBGL-SS	204.33	318.80	0.505	0.1265	0.415
LSGL	44.635	188.81	0.544	0	0.479
SRRR	242.67	193.64	0.594	0	0.587

Experiment 3: $n = 100, p = 500, q = 3$. 10 active predictors

Method	$\mathbf{MSE_{est}}$	$\mathbf{MSE_{pred}}$	**FDR**	**FNR**	**MP**
MBSP	**0.185**	**64.14**	**0.048**	0	**0.0011**
MBGL-SS	1.327	155.51	0.483	0.0005	0.092
LSGL	0.2305	72.894	0.849	0	0.117
SRRR	0.9841	49.428	0.688	0	0.104

Table: Simulation results for MBSP-TPBN, compared with thee other methods, averaged across 100 replications.

Yeast Cell Cycle Data Analysis

Transcription factors (TFs) are sequence-specific DNA binding proteins which regulate the transcription of genes from DNA to mRNA by binding specific DNA sequences. We want to know which TFs are significant.

In this yeast cell cycle data set (first studied by Chun and Keles (2010)):

- mRNA levels are measured at 18 time points seven minutes apart (every 7 minutes for a duration of 119 minutes).
- The 542×18 response matrix $\mathbf{Y}$ consists of 542 cell-cycle-regulated genes from an α factor arrested method, with columns corresponding to the mRNA levels at the 18 distinct time points. The 542×106 design matrix $\mathbf{X}$ consists of the binding information of a total of 106 TFs.

We fit the MBSP model to this data set. We assess its predictive performance using 5-fold cross validation and perform variable selection from the 106 TFs.

Yeast Cell Cycle Data Analysis

Method	Number of Proteins Selected	MSPE
MBSP	10	18.491
MBGL-SS	7	20.093
LSGL	4	22.819
SRRR	44	18.204

Table: Results for analysis of the yeast cell cycle data set. The MSPE has been scaled by a factor of 100. In particular, all four models selected the three TFs, ACE2, SWI5, and SWI6 as significant.

The SRRR method has the lowest MSPE but it recovers a *non*-parsimonious model. In contrast, MBSP has good predictive performance *and* recovers a parsimonious model.

Yeast Cell Cycle Data Analysis

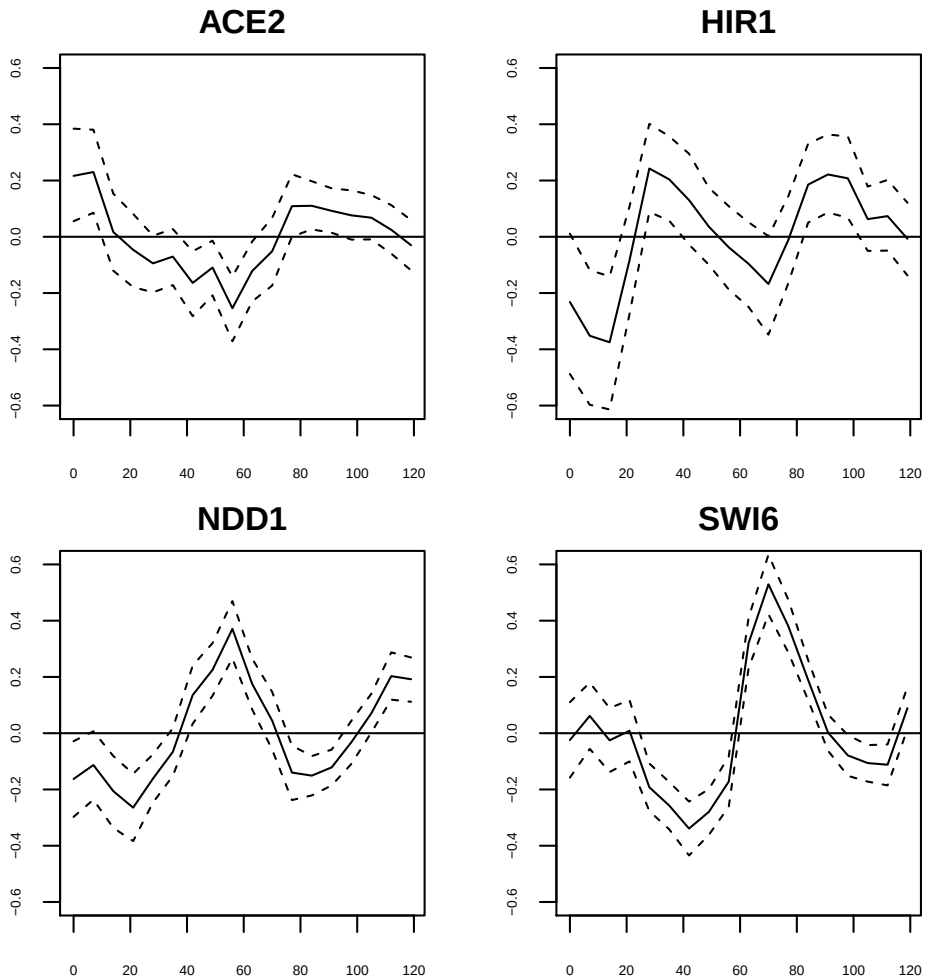

Figure: Plots of the estimates and 95% credible bands for four of the 10 TFs that were deemed as significant by the MBSP-TPBN model. The x-axis indicates time (minutes) and the y-axis indicates the estimated coefficients.

Summary of MBSP Model

We have introduced a new Bayesian approach known as the Multivariate Bayesian model with Shrinkage Priors (MBSP) for the multivariate linear regression model, $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$.

- Our model produces a row-sparse estimate of the $p \times q$ matrix, $\mathbf{B}$, allowing for sparse estimation and variable selection from the p variables.
- Our model can consistently estimate $\mathbf{B}$ even when $p \gg n$ and p grows at nearly exponential rate with n (i.e. $p = O(e^{n^d}), 0 < d < 1$.)
- A wide variety of polynomial-tailed shrinkage priors may be used, so our model and our theoretical results are quite general.
- We illustrated practical application of our model with the three parameter beta normal family (MBSP-TPBN), using the horseshoe prior as a special case.

Future Work

Open problems:

- Theoretical investigation of MBSP (and Bayesian multivariate regression models in general) when $q \to \infty$ and when $\mathbf{\Sigma}$ is treated as unknown.
- Moving beyond consistency, deriving a particular contraction rate of the MBSP's posterior around $\mathbf{B}_0$.
- Applying polynomial-tailed priors to reduced rank regression and partial least squares regression.

Pre-print of Paper

A pre-print of the paper for this presentation is available at:
https://arxiv.org/abs/1711.07635

Accepted pending minor revision at *Journal of Multivariate Analysis*.

References

- Armagan, A., Clyde, M., and Dunson, D.B. (2011) "Generalized Beta Mixtures of Gaussians." *Advances in Neural Information Processing Systems 24*, 523-531.
- Armagan, A., Dunson, D.B., Lee, J., Bajwa, W., and Strawn, N. (2013) "Posterior Consistency in Linear Models Under Shrinkage Priors." *Biometrika*, 100(4): 1011-1018.
- Brown, P.J., Vannucci, M., and Fearn, T. (1998) "Multivariate Bayesian Variable Selection and Prediction." *Journal of the Royal Statistical Society: Series B*, 60(3): 627-641.
- Carvalho, C.M., Polson, N.G., and Scott, J.G. (2010) "The Horseshoe Estimator for Sparse Signals." *Biometrika*, 97(2):465-480.

References


Chen, L. and Huang, J.Z. (2012) "Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection." *Journal of the American Statistical Association*, 107(500): 1533-1545.


Li, Y., Nan, B., and Zhu, J. (2015) "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure." *Biometrics*, 71(2): 354-363.


Liquet, B., Mengersen, K., Pettitt, A.N., and Sutton, M. (2017) "Bayesian Variable Selection Regression of Multivariate Responses for Group Data." *Bayesian Analysis* 12(4): 1039-1067.


Tang, X., Xu, X., Ghosh, M., and Ghosh, P. (2017) "Bayesian Variable Selection and Estimation Based on Global-Local Shrinkage Priors." *Sankhya A*.

Questions?